УДК 519.7

Н.В. Борисова, О.В. Канищева, О.И. Король

ИДЕНТИФИКАЦИЯ МОРФОЛОГИЧЕСКИХ ХАРАКТЕРИСТИК НОВОГО СЛОВА ПРИ СОСТАВЛЕНИИ ЭЛЕКТРОННЫХ СЛОВАРЕЙ

Введение. Стремительные темпы развития современного общества напрямую отражаются в языке – возникают новые слова, термины и выражения, изменяется смысл старых слов. Традиционные "бумажные" словари практически не успевают реагировать на эти изменения, потому что они слишком долго готовятся к печати. Поэтому на смену им приходят словари электронные – компьютерные программы, позволяющие оперативно получить необходимую информацию. Часто считают, что электронный словарь – это просто электронная версия некоторого "бумажного" словаря, снабженная удобным пользовательским интерфейсом и машинными средствами поиска. Однако, мы разделяем иную точку зрения: электронный словарь – это новая форма словаря, которая позволяет устранить многие неизбежные недостатки "бумажной" лексикографии и тем самым поднять ее на качественно новый уровень. Хотя разработки в области автоматизации создания электронных словарей ведутся довольно активно, процесс составления подобных словарей все еще является достаточно трудоемким.

Таким образом, задача идентификации новых слов с целью пополнения ими электронных словарей является актуальной и практически значимой.

Постановка задачи. Нашей задачей является автоматизировать процедуру склонения и спряжения имен существительных, при минимальном привлечении пользователя для идентификации морфологических характеристик нового слова при составлении электронных словарей.

В своей работе мы использовали аппарат алгебры конечных предикатов, разработанный школой Ю.П. Шабанова-Кушнаренко [1, 2]. Представим математическое описание склонения на примере имен существительных.

Математическое описание склонения имен существительных. Поскольку наполнением большинства словарей являются имена существительные, в качестве примера мы выбрали именно эту часть речи.

Предложенная нами математическая модель связывает буквы окончания и грамматические признаки словоформ имен существительных с этими окончаниями. Важно отметить, что модель описывает далеко не все явления, имеющие место при склонении имен существительных, а только ту часть из них, которая подчиняется достаточно компактной группе правил.

Словоформы субстантивного склонения, окончания и грамматические признаки которых удовлетворяют морфологическому отношению, которое приведено в предлагаемой модели, мы назовем основными, все остальные словоформы относим к классу неосновных. В дальнейшем под словоформой будем понимать не только соответствующий ей текст, но и относящейся к ней набор грамматических признаков.

Рассматривать одновременно все буквы словоформ мы не имеем возможности, так как это привело бы к чрезмерно обширной задаче. Поэтому ограничимся пока лишь одной буквой, а именно *первой буквой окончания* в словоформах. Будем считать, что первая буква окончания – всегда гласная. Как показывает просмотр всевозможных окончаний словоформ, первой буквой окончания может быть любая гласная, кроме буквы э. В некоторых словоформах окончание отсутствует вовсе (*стол, мест*) или же в нем на первом месте нет гласной буквы (*сарай, пень, детьми*). В этом случае условимся считать, что на месте первой буквы окончания находится знак пробела *. Будем обозначать переменную первую букву окончания символом y_1. С учетом принятых соглашений область изменения для переменной y_1 можно задать следующим уравнением алгебры конечных предикатов:

$$y_1^{а} \vee y_1^{е} \vee y_1^{ё} \vee y_1^{и} \vee y_1^{о} \vee y_1^{у} \vee y_1^{ы} \vee y_1^{ю} \vee y_1^{я} \vee y_1^{*} = 1. \qquad (1)$$

Несмотря на то, что мы ограничились рассмотрением только одной буквы в словоформах, задача все еще остается необозримо обширной, и мы вынуждены подвергнуть ее дальнейшему сужению. С этой целью воспользуемся теоремой о дизъюнктивном разложении предиката. Как было показано в работе [3], объектом математического описания в морфологии служит морфологическое отношение, связывающее фрагмент словоформы (в данном случае – первую букву ее окончания) со смыслом этого фрагмента. Смысл первой буквы окончания y_1 представим в виде набора $(x_1, x_2, \ldots, x_n)$ с буквенными компонентами $(x_1, x_2, \ldots, x_n)$. Нашей задачей является описание средствами алгебры конечных предикатов морфологического отношения:

$$L(x_1, x_2, \ldots, x_n, y_1) = 1. \tag{2}$$

Разложим предикат L, стоящий в левой части канонического уравнения (2), (назовем его *морфологическим предикатом*) по одной из его переменных, **например** по переменной x_1:

$$\begin{aligned} &L(x_1, x_2, \ldots, x_n, y_1) = x_1^{a_1}(a_1, x_2, \ldots, x_n, y_1) \vee \\ &x_1^{a_2}(a_2, x_2, \ldots, x_n, y_1) \vee \ldots \vee x_1^{a_k}(a_k, x_2, \ldots, x_n, y_1). \end{aligned} \tag{3}$$

При разложении принято, что переменная пробегает значения $a_1, a_2, \ldots, a_k$. Принимая поочередно для переменной x_1 значения $a_1, a_2, \ldots, a_k$ получаем k частных случаев морфологического предиката:

$$\begin{aligned} &L_1(x_2, \ldots, x_n, y_1) = L(a_1, x_2, \ldots, x_n, y_1), \\ &L_2(x_2, \ldots, x_n, y_1) = L(a_2, x_2, \ldots, x_n, y_1), \\ &\ldots\ldots\ldots\ldots \\ &L_k(x_2, \ldots, x_n, y_1) = L(a_k, x_2, \ldots, x_n, y_1). \end{aligned} \tag{4}$$

Отношения

$$\begin{aligned} &L_1(x_2, \ldots, x_n, y_1) = 1, \\ &L_2(x_2, \ldots, x_n, y_1) = 1, \\ &\ldots\ldots\ldots\ldots \\ &L_k(x_2, \ldots, x_n, y_1) = 1 \end{aligned} \tag{5}$$

зависят от меньшего числа переменных, чем исходное морфологическое отношение (2), поэтому каждое из них изучать проще. После того как будут порознь изучены отношения (5), уже не составит большого труда по формуле (3) собрать из соответствующих им предикатов $L_1, L_2, \ldots, L_k$ морфологический предикат:

$$\begin{aligned} &L(x_1, x_2, \ldots, x_n, y_1) = x_1^{a_1} L_1(a_1, x_2, \ldots, x_n, y_1) \vee \\ &x_1^{a_2} L_2(a_2, x_2, \ldots, x_n, y_1) \vee \ldots \vee x_1^{a_k} L_k(a_k, x_2, \ldots, x_n, y_1) \end{aligned} \tag{6}$$

и записать с его помощью морфологическое отношение (2). Если же и предикаты $L_1, L_2, \ldots, L_k$ окажутся слишком сложными для непосредственного математического описания, то к ним тоже можно применить теорему о разложении и т.д. до тех пор, пока не придем к достаточно простым предикатам.

Проведем теперь описанным методом сужение нашего объекта математического описания. С этой целью введем пять грамматических признаков (компонентов смысла) и зафиксируем их значения. Прежде всего введем переменную ξ_1, называемую нами *видом морфа* со значениями о – окончание, н – не окончание (корень, суффикс и др.). Принимаем ξ_1=о, этим выбором мы ограничиваем задачу рамками словоизменения. Если бы мы приняли ξ_1=н, то вышли в область словообразования. Для всех шести вводимых признаков мы предусматриваем только по два значения. В качестве первого значения выбираем то, на котором решили остановиться, ко второму же значению относим все остальные традиционно вводимые в грамматике градации признака. Нам нет надобности проводить детализацию второго значения, поскольку его мы все равно пока не собираемся рассматривать. Такую детализацию можно провести впоследствии, когда будет рассматриваться объект, вводимый вторым значением признака.

Вторым компонентом смысла вводим *номер буквы окончания* ξ_2 со значениями п – первый, н – не первый (второй, третий). Принимаем ξ_2=п, тем самым ограничивая себя лишь рассмотрением первой буквы окончания. Далее вводим признак ξ_3 – *тип словоизменения* со значениями с – склонение, н – не склонение. Принимая ξ_3=с, мы при этом ограничиваем нашу задачу изучением процесса склонения имен. Если же мы взяли бы ξ_3=н, то вступили в область спряжения глаголов. Вводим переменную ξ_4 – *тип склонения* со значениями с – субстантивный и н – не субстантивный (адъективный и др.). Принимаем ξ_4=с, ограничиваясь изучением склонения имен существительных. Принятие второго значения н переменной ξ_4 привело бы нас к изучению склонения имен прилагательных и некоторых других классов

имен. Наконец, вводим признак ξ_5 – *вид словоформы* со значениями о – основной и н – не основной. Принимаем ξ_5=о. Введение признака ξ_5 и принимаемая его фиксация продиктованы желанием не рассматривать на первых порах сравнительно немногочисленные словоформы, отклоняющиеся от единообразной основной массы словоформ.

Введение признаков $\xi_1 \div \xi_5$ и фиксация их значений позволяет нам, используя теорему о разложении, существенно сузить объект математического описания и оставить в поле зрения только словоформы, получаемые при склонении некоторых имен существительных. Словоформы субстантивного и адъективного склонения выпадают из рассмотрения.

Теперь группа словоформ, выбранная нами для математического описания, достаточно четко очерчена, и нам остается сформировать полный (и несократимый) набор переменных признаков (компонентов смысла) для первой буквы окончаний словоформ этой группы. Мы останавливаемся на следующих восьми признаках: x_1 – *падеж* со значениями *и* – именительный, *р* – родительный, *д* – дательный, *в* – винительный, *т* – творительный, *п* – предложный; x_2 – *род* со значениями *м* – мужской, *ж* – женский, *с* – средний; x_3 – *число* со значениями *е* – единственное, *м* – множественное; x_4 – *признак одушевленности* со значениями *о* – одушевленный, *н* – неодушевленный; x_5 – *признак ударности* основы со значениями *у* – ударный, *б* – безударный; x_6 – *требование простановки знака йотирования* над буквой ё со значением *с* – ставить, *н* – не ставить; x_7 – *последняя буква основы* словоформы со значениями *а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ы, ю, я*; x_8 – *вид основы* словоформы со значениями *т* – твердый, *м* – мягкий.

Признаки x_1, x_2, x_3, x_4 полностью совпадают с введенными нормативной грамматикой понятиями падежа, рода, числа и признака одушевленности и в комментариях не нуждаются. Значение признака ударности понимаем равным x_5=*у*, если основа словоформы ударная, и x_5=*б* – в противном случае. Значение требования простановки знака ¨ принимаем равным x_6=*с*, если две точки над ё в первой букве окончания должны быть проставлены, и x_6=*н* – в случае, если буква ё должна быть записана в виде е. Под последней буквой основы x_7 понимаем букву словоформы, предшествующую окончанию, **например**, в словоформе *столом* x_7=*л*. Под это определение не попадает мягкий знак: если он стоит впереди окончания, то в качестве последней буквы основы принимаем букву, предшествующую мягкому знаку, **например**, в словоформах *карась, полесье* полагаем x_7=*с*. Как показывает просмотр орфографического словаря, в роли последней буквы основы в избранном нами классе словоформ буквы *ё, ъ, э* не встречаются. Если бы мы ограничились только признаками $x_1 \div x_7$, то однозначное управление первой буквой окончания не было бы достигнуто. При различной фиксации значений этих признаков в некоторых случаях получаются двойные значения для переменной y_1: *а – я* (*топора, словаря*), *у – ю* (*топору, словарю*), *о – ё* (*топором, словарём*), *ы – и* (*топоры, словари*), *о – е* (*топоров, словарей*) [4]. Это колебание значений переменной y_1 будем трактовать как результат влияния вида основы словоформы x_8. Полагаем, что при x_8=*т* имеет место первое значение переменной y_1, при x_8=м – второе.

Области изменения введенных переменных формально зададим следующими уравнениями:

$$x_1^{и} \vee x_1^{р} \vee x_1^{д} \vee x_1^{в} \vee x_1^{т} \vee x_1^{п} = 1, \quad x_2^{м} \vee x_2^{ж} \vee x_2^{с} = 1, \quad x_3^{е} \vee x_3^{м} = 1, \quad x_4^{о} \vee x_4^{н} = 1, \quad x_5^{у} \vee x_5^{б} = 1, \quad x_6^{с} \vee x_6^{н} = 1,$$

$$x_7^{а} \vee x_7^{б} \vee x_7^{в} \vee x_7^{г} \vee x_7^{д} \vee x_7^{е} \vee x_7^{ж} \vee x_7^{з} \vee x_7^{и} \vee x_7^{й} \vee x_7^{к} \vee x_7^{л} \vee x_7^{м} \vee x_7^{н} \vee x_7^{о} \vee x_7^{п} \vee x_7^{р} \vee x_7^{с} \vee x_7^{т} \vee x_7^{у} \vee x_7^{ф} \vee x_7^{х} \vee$$

$$\vee x_7^{ц} \vee x_7^{ч} \vee x_7^{ш} \vee x_7^{щ} \vee x_7^{ы} \vee x_7^{ю} \vee x_7^{я} = 1 \quad , \; x_8^{т} \vee x_8^{м} = 1 .$$

Введенных признаков достаточно для однозначного задания их значениями первой буквы окончания любой основной словоформы субстантивного склонения. Нетрудно убедиться в том, что введенный набор признаков не только полон, но и несократим. В самом деле, исключая из него падеж, получаем неоднозначность типа, **например**, *ножом, ножа* (y_1=*о, а*); исключая род – неоднозначность типа *сосны*, *окна* (y_1=*ы, а*); число – *ножа, ножей* (y_1=*а, е*); признак одушевленности – *вижу портфель, вижу жителя* (y_1=**, е*); признак ударности основы – *недель, тлей* (y_1=**, е*); требование простановки знака ¨ – *остриё, острие* (y_1=*ё, е*); последнюю букву основы – *о солнце, о здании* (y_1=*е, и*); вид основы – *топоры, словари* (y_1=*ы, и*).

Итак, мы полностью завершили постановку задачи. Она состоит в том, чтобы средствами алгебры конечных предикатов записать морфологическое отношение:

$$L(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, y_1) = 1, \tag{7}$$

связывающее первую букву окончания y_1 основных словоформ субстантивного склонения с грамматическими признаками $x_1 \div x_8$. В силу полноты набора $(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$ уравнение (7) задает некоторую функцию:

$$y = F(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) \tag{8}$$

Дальнейшее усовершенствование предложенной модели будет состоять в представлении предикатов F_σ в виде формул алгебры конечных предикатов, где $\sigma \in \{a, e, ё, u, o, y, ы, ю, я, *\}$, задающие всевозможные узнавания y_1^σ для переменной y_1:

$$y_1^\sigma = F_\sigma(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) \quad (9)$$

Заключение. В данной работе для идентификации морфологических характеристик новых слов при составлении электронных словарей предлагается использовать аппарат алгебры конечных предикатов и предикатных операций.

Недостатком данной модели является ее незавершенность, т.е. в модели все грамматические признаки выступают в роли независимых переменных. Однако в действительности между грамматическими признаками имеют место другие связи, существенно ограничивающие число допустимых наборов значений грамматических признаков для словоформ тех или иных конкретных слов. Эти связи в модели не отражены. В дальнейшем предполагается решение и этой задачи, а, следовательно, усовершенствования предложенной модели.

ЛИТЕРАТУРА:

1. Шабанов-Кушнаренко Ю.П. Теория интеллекта: Проблемы и перспективы. Ю.П. Шабанов – Кушнаренко – Х.: Вища шк., 1987. – 158 с.
2. Шабанов-Кушнаренко Ю.П., Бондаренко М.Ф. Математическая модель склонения непритяжательных имен прилагательных // Науч.-техн. информация. Сер. 2. – 1979. – № 6. – С. 10-13.
3. Боярский К., Каневский Е. Методика пополнения компьютерного словаря, используемого при разметке корпусов текстов // Прикладна лінгвістика та лігвістичні технології: MegaLing – 2007: зб. наук. пр. / НАН України, Укр. мовно-інформ. фонд. – К.: Довіра, 2008. – С. 85-93.
4. Зализняк А.А. Грамматический словарь русского языка. – М.: Рус. яз., 1977. – С. 39-53. – 400 с.

БОРИСОВА Наталья Владимировна – аспирант кафедры интеллектуальных компьютерных систем Национального технического университета "Харьковский политехнический институт".

Научные интересы: автоматизированные библиотечные системы, создание электронных словарей.

КАНИЩЕВА Ольга Валерьевна – старший преподаватель кафедры интеллектуальных компьютерных систем Национального технического университета "Харьковский политехнический институт".

Научные интересы: автоматизированная обработка естественного языка, информационно-поисковые системы, автоматизированные библиотечные системы.

КОРОЛЬ Ольга Игоревна – ассистент кафедры информатики и интеллектуальной собственности Национального технического университета "Харьковский политехнический институт".

Научные интересы: интеллектуальный анализ данных, защита интеллектуальной собственности, патентно-конъюнктурные исследования.